

(19)



Europäisches Patentamt  
European Patent Office  
Office européen des brevets

(11) Veröffentlichungsnummer:

(11) Publication number:

(11) Numéro de publication:

**0 857 383**

Internationale Anmeldung veröffentlicht durch die  
Weltorganisation für geistiges Eigentum unter der Nummer:

**WO 97/16004** (art.158 des EPÜ).

International application published by the World  
Intellectual Property Organisation under number:

**WO 97/16004** (art.158 of the EPC).

Demande internationale publiée par l'Organisation  
Mondiale de la Propriété sous le numéro:

**WO 97/16004** (art.158 de la CBE).

**THIS PAGE BLANK (USPTO)**

PCT

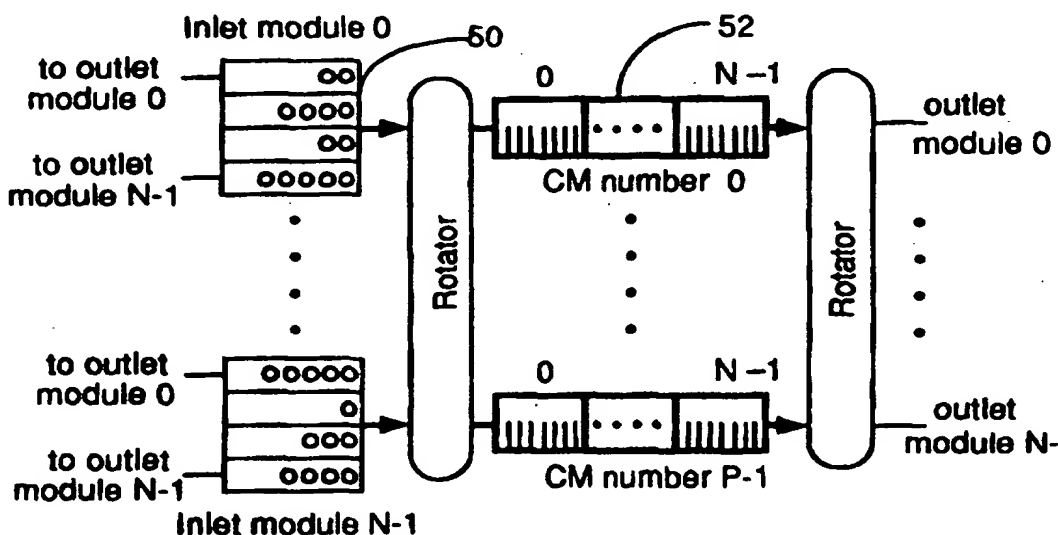
WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification <sup>6</sup> : H04L 12/56	A1	(11) International Publication Number: WO 97/16004
		(43) International Publication Date: 1 May 1997 (01.05.97)
(21) International Application Number: PCT/CA96/00673 (22) International Filing Date: 9 October 1996 (09.10.96) (30) Priority Data: 08/548,716 26 October 1995 (26.10.95) US (71) Applicant: NORTHERN TELECOM LIMITED [CA/CA]; World Trade Center of Montreal, 8th floor, 380 St. Antoine Street West, Montreal, Quebec H2Y 3Y4 (CA). (72) Inventors: BESHAI, Maged, E.; 70 Trailway Circle, Stittsville, Ontario K2S 1E2 (CA). MUNTER, Ernst, A.; 4 Nanook Crescent, Kanata, Ontario K2L 2A7 (CA). (74) Agent: TOYOOKA, Yoshiharu; Northern Telecom Limited, Patent Dept., Station "C", P.O. Box 3511, Ottawa, Ontario K1Y 4H7 (CA).		(81) Designated States: CA, JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).  Published With international search report.

(54) Title: HIGH CAPACITY ATM SWITCH



(57) Abstract

An ATM switch architecture expandable to multi-terabits/s uses data transfer in a heterogeneous burst of a constant length. It employs rotators connecting stages in a three-stage switch configuration. In one embodiment, the cells are sorted at ingress and a matching process is performed between the first and middle stages. The switch is simple to control and has high performance at both the call and cell levels. It also meets the basic requirements that cells be delivered in the proper order, and that the rate of any individual connection be as high as the inlet-port rate. With a small internal expansion, the switch is non-blocking in the sense that any bit-rate acceptable to both the inlet and outlet ports will be guaranteed a path through the core. This feature is particularly useful in services which may require frequent bit-rate change during the connection time.

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgyzstan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic			SE	Sweden
CG	Congo	KR	Republic of Korea	SG	Singapore
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LR	Liberia	SZ	Swaziland
CS	Czechoslovakia	LT	Lithuania	TD	Chad
CZ	Czech Republic	LU	Luxembourg	TG	Togo
DE	Germany	LV	Latvia	TJ	Tajikistan
DK	Denmark	MC	Monaco	TT	Trinidad and Tobago
EE	Estonia	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	UG	Uganda
FI	Finland	ML	Mali	US	United States of America
FR	France	MN	Mongolia	UZ	Uzbekistan
GA	Gabon	MR	Mauritania	VN	Viet Nam

## HIGH CAPACITY ATM SWITCH

### Technical Field and Industrial Applicability

The invention generally relates to ATM switches. In particular,  
5 it is directed to high capacity ATM switches which use rotators and  
common memory modules.

### Background Art

Traditional ATM switches are primarily cell-synchronous. The  
10 two most popular configurations used in large-scale switching nodes  
are the buffer-space-buffer and the three-buffer-stage networks. Fig. 1  
shows a typical buffer-space-buffer network. An  $N \times N$  single-stage space  
switch interconnects  $N$  asynchronous multiplexers to  $N$  asynchronous  
demultiplexers. Priority queuing may be provided at the inlet modules  
15 in order to control the quality-of-service (QOS) of traffic streams  
belonging to different classes. The inlet multiplexers and outlet  
demultiplexers may be paired to form a folded architecture with intra-  
module switching. With asynchronous multiplexing at inlet, this  
configuration requires a fast mechanism for contention resolution. In  
20 the classical buffer-space-buffer architecture, arbitration to resolve  
multiple simultaneous demand for a given outlet is done on a cell-by-  
cell basis, requiring a fast mechanism.

Figure 2 depicts a known plain three-buffer-stage configuration.  
This configuration does not have a contention problem, thanks to the  
25 extra buffering stage, but has some capacity limitations. In the  
architecture of Figure 2, each component is an  $n \times n$  common-memory  
(CM) or output-buffered (OB) switch;  $n$  is typically 16 or so and there  
are  $P$  middle modules. With  $P=n$ , the total capacity is limited to  $n^2$   
times the link speed  $r$ . In the folded architecture which will be  
30 described below, the capacity limit is  $(1/2) n^2$  times the link speed  $r$ .  
The cells of a given connection, between different outer modules, must  
be routed through the same intermediate switching module in order to  
guarantee proper cell order. Priority service can be implemented at one  
or more stages.

35 As mentioned above, the cell-synchronous switches have  
capacity limitations. U.S. Patent No. 5,475,679 issued on December 12,  
1995 (Munter), describes a switch architecture suitable for very high-

speed networks. The design was guided by two main principles. The first is to transfer multiple cells, padded by a reasonable guard time, to circumvent the high speed cell synchronization problem. The second is to sort the incoming cells at ingress to facilitate internal routing and congestion control within the switch. The multiple cells, hereafter called bursts, must belong to the same egress port, and the burst length could vary significantly from one cell to a hundred cells or so. The bursts are transferred directly from inlet to outlet through an optical space switch and a central controller is used to realize a collision-free transfer. As such, the switch capacity is limited mainly by the speed of the controller.

Specifically speaking, the classical buffer-space-buffer architecture has a single input buffer (perhaps per class), and the destination information is only stored in the cell headers. As shown in Figure 3, in the architecture of the above co-pending application the cells are sorted according to destination, thus facilitating the contention resolution task. In Figure 3, the common buffer of each inlet module (inlet to space switch) is divided into a number of variable length sections. The number of sections is N or less, depending on the number of inlet modules. Priority service can be implemented by a further subdivision of each section according to the number of classes per destination. The inter-module payload transfer is based on requests and grants. An inlet module which has cells to send to an outlet module must signal its intention to do so. The control system decides the time of the load transfer and the number of cells in each transaction. The load is transferred in the form of homogeneous bursts; a homogeneous burst contains cells of the same destination as depicted in Figure 4. The idle slots shown in Figure 4 represent the inter-burst guard time. This gives rise to two possibilities: a centralized-control protocol, or a protocol based on distributed control. The architecture in the copending application is based on central control. An inlet module makes a request by simply indicating the required destination and the number of cells in the current load. This information is sent through a control bus accessed periodically (or by any other suitable discipline) by the central controller. Cell order is naturally preserved since requests are processed one at a time. The highest individual connection rate equals the inlet-port rate, for example 600Mb/s or so. The capacity of

the switch is limited primarily by the controller speed. Even with a dedicated processor per inlet controlling the traffic flow, excessive delays would occur when the number of destinations is large. With a relatively small number of inlet (outlet) modules,  $N=16$  for example, the cell delay performance is excellent. Also, the inlet-buffer requirement is quite modest at relatively high traffic loads. Thus, an infinitesimal cell-loss is realizable with a reasonable buffer size.

In U. S. Patent No. 5,168,492 (Beshai et al), issued December 1, 1992, rotating access ATM/STM packet switches are described which are functionally equivalent to the classical buffer-space-buffer architecture. In the basic embodiment, it uses middle packet buffers with a rotator (commutator) at its input and output.

The use of burst transfer, optical rotators, and distributed control facilitates the construction of high capacity switches using lower capacity modules. According to the present invention, a significant capacity increase can be realized if rotators are used and several controllers operate simultaneously on non-overlapping inlet-outlet pairs. This can be achieved in a simple manner if the condition that a burst must contain cells of the same destination is relaxed, and if the bursts are of equal size. In the architecture of Figure 3, the bursts are homogeneous (i.e., all the burst cells have the same destination) and of variable length, as shown in Figure 4. The control can be enhanced if the bursts are heterogeneous and of equal size as shown in Figure 5. A heterogeneous burst may contain cells of different destinations. The invention therefore uses the concepts described in the above-referenced copending patent application and the rotating-access idea of U.S. Patent No. 5,168,492 to construct a switch with an ultimate capacity of several tera bits/s. The maximum connection rate, which is the permissible rate for a single user, is the inlet port speed.

### Objects of the Invention

It is an object of the invention to provide a high capacity ATM switch which employs rotators in a three-stage configuration and transfers data in heterogeneous bursts of a predetermined length.

It is another object of the invention to provide a method of switching data in heterogeneous bursts of a predetermined length.

It is a further object of the invention to provide a high capacity ATM switch which uses matching of cells between inlet buffers and middle buffers.

5 It is yet another object of the invention to provide a method of switching data in heterogeneous bursts of a predetermined length which includes a step of matching cells between the inlet and middle stages.

It is still another object of the invention to provide a high capacity ATM switch which is internally non-blocking.

10

#### Disclosure of the Invention

Briefly stated, according to one aspect, the invention relates to a high capacity ATM switching system for switching data in a burst of a predetermined number of cells among N inlet modules and M outlet modules in each successive access time, M and N being positive integers. The switching system comprises the N inlet modules having buffers, each buffer dedicated to each of the outlet modules, for storing cells according to the destination outlet modules of the cells in respective buffers and P common memories, P being a positive integer, 15 each common memory having M memory sections, each of which is able to hold at least said predetermined number of cells and is dedicated to each outlet module. The switching system further includes an inlet rotator for cyclically connecting in each access time the N inlet modules and P common memories so that respective cells are transferred from the N inlet modules and stored in respective 25 sections according to the destination outlet module of each cell, and an outlet rotator for cyclically connecting in each access time the P common memories and M outlet modules so that respective outlet modules are connected to respective memory sections for reading out cells contained therein. 30

According to another aspect, the invention is directed to a method of switching data in a burst of a predetermined number of cells among N inlet modules and M outlet modules in each successive access time, M and N being positive integers. The method comprises 35 steps of each of the N inlet modules storing cells in separate buffers according to the destination outlet modules of the cells and cyclically connecting the N inlet modules and P common memories, P being a



positive integer. The method further includes steps of transferring in each access time the burst of the predetermined number of cells from one of the N inlet modules to respective memory sections of one of the common memories according to the destination outlet modules of the cells, and cyclically connecting the common memories and M outlet modules so that respective outlet modules are connected to the respective memory section for reading out cells contained therein.

#### Brief Description of the Drawings

- Figure 1 is a known buffer-space-buffer switch;  
Figure 2 is a known three-stage switch;  
Figure 3 shows a buffer-space switch with inlet sorting and burst transfer;  
Figure 4 shows variable-length homogeneous bursts;  
Figure 5 shows constant-length heterogeneous bursts;  
Figure 6 is a three-stage rotator-linked switch according to one embodiment of the invention;  
Figure 7 is a three-stage switch with inlet sorting, burst transfer, and distributed control according to another embodiment of the invention;  
Figure 8 shows the operation of rotating-access to middle CMs according to the invention;  
Figure 9 illustrates the matching process of the invention;  
Figure 10 shows the simultaneous matching process;  
Figure 11 shows yet a further embodiment of the invention which uses a ring configuration;  
Figure 12 is a switch system in folded architecture;  
Figure 13 is a common memory multiplexer/demultiplexer switch;  
Figure 14 is a  $k^2 \times k^2$  rotator requiring  $2k$  units of smaller  $k \times k$  rotators;  
Figure 15 is a  $k^3 \times k^3$  rotator requiring  $3k^2$  units of smaller  $k \times k$  rotators;  
Figure 16 is a graph showing inlet-buffer occupancy distribution;  
and  
Figures 17, 18 and 19 are graphs showing cell-delay complementary functions under different conditions.

### Mode(s) of Carrying Out the Invention

Figure 6 shows schematically a switching architecture according to one embodiment of the invention. In this embodiment,  $N$  inlet modules 30 and  $N$  outlet modules 32 are linked by two optical rotators 34, 36 and  $P$  middle modules 38. Each inlet module receives data from  $n$  inputs 40 and sends multiplexed data to rotator 34 through a serial link 42. Each outlet module accepts multiplexed data from rotator 36 and demultiplexes them to  $n$  outputs 44.  $N$ ,  $P$  and  $n$  are any positive integer numbers. It is also possible to have different numbers of inlet and outlet modules. In this embodiment, the inlet, middle and outlet modules are made of several common memory modules and designated  $CM_0$ - $CM_{N-1}$  and  $CM_0$ - $CM_{P-1}$ . Each rotator is a  $k \times k$  rotator,  $k > 1$  which is a periodic selector, equitably connecting each of its inputs to each of its outputs. In other words, it is a counter-driven  $k \times k$  selector. It functions as  $k$  parallel sets of  $k$  serial links (a total of  $k^2$  links). With identical inlets of speed, e.g.,  $\chi$  b/s each, the speed of each link is  $\chi/k$  b/s. These links are hereafter called "virtual links". They are called virtual because they are reconfigurable. A virtual link connects an outer CM to a middle CM during a fixed interval of several time slots (a time slot is the cell duration). This interval is called the "access time", denoted  $\Delta$ .

A burst of cells of possibly different destinations is transferred from an inlet CM to a middle CM per access time. The highest individual-connection rate is  $r \cdot (n/P)$ , where  $n$  is the number of external ports per CM,  $P$  is the number of middle CMs and  $r$  is the speed of an external port. This configuration works almost exactly like the three-stage switch shown in Figure 2, the only difference being that the links from a given outer module to the set of middle modules are fast and intermittent in this architecture as compared to slower and continuous in the standard architecture of Figure 2; hence the need for burst transfer.

This architecture does not require control communications between the stages. The capacity is virtually unlimited. The overall cell delay variance, however, may be unacceptable for CBR (constant bit rate) and other delay-sensitive traffic. This problem can be solved by appropriate path selection for the virtual circuits at the call-admission stage and by providing priority classification, at least at the middle CMs.

Basically, the internal-routing mechanism should distribute the delay-sensitive traffic equitably among the middle CMs, where they are given high transfer priority to the output modules. There are different traffic classifications such as the CBR, the VBR (variable bit rate), and the ABR (available bit rate). An individual connection, regardless of classification, must use the same middle CM in order to maintain proper cell sequence. This requirement limits the highest connection rate per user to  $r \cdot n / P$ . For example, if  $r = 620$  Mb/s,  $n = 16$ , and  $P = 256$ , the capacity of the switch is approximately 2.5 Tb/s but the highest connection rate is less than 40Mb/s ( $620 \times 16 / 256$ ). Thus, the price of high capacity is a reduced upper bound of individual connection rates.

The same high capacity, but with an individual connection rate as high as the external port speed, is realized with further controls as will be described below in connection with a further embodiment of the invention.

Figure 7 shows such an embodiment which uses a matching process. During a rotator cycle, each inlet module 50 visits each middle CM 52. The access time, denoted  $\Delta$  (slots), during each visit is fixed. At a rotator port speed of 10Gb/s, for example, a value of  $\Delta = 16$  corresponds to about 0.7  $\mu$ sec. During each access time, a number of cells belonging to one or more outlet modules is transferred. A guard time of one or two cells may be needed within each access time. The productive time of the access time is hereafter called the duty cycle, and the number of cells per duty cycle is denoted  $D$ . A control array within each middle CM stores the number of cells destined to each of the  $N$  outlet modules. There are  $N$  inlet and outlet modules in this embodiment but unequal numbers are possible. During each access time, each outlet module reads the cells destined to it and resets the corresponding entry of the control array to zero. The maximum number of cells read per access time is  $D$  (14, for example, if  $\Delta$  is chosen to be 16 cells and a guard time of two cells is used). The admission of cells to the middle CMs is based on a matching process. The capacity is dependent on the size of the rotator and is virtually unlimited. There is a constant delay from each inlet to each outlet. This delay varies from one access time (of the order of 1  $\mu$ sec) to  $N$  access times, but is constant for the same inlet-outlet pair. For example, with 16 OC12 (optical carrier, about 620 Mb/s) ports per inlet CM module, a time slot (ATM cell duration) at the

optical rotator port is about 40 nsec. Selecting an access time of 16 slots (about 0.7  $\mu$ sec), the worst constant delay in a large switch with 256 middle CM modules (2.5 Tb/s capacity) is less than 200  $\mu$ sec.

During each access time (of 16 slots duration for example), each  
5 inlet module transfers a burst of cells to a middle module. The number of transferred cells is limited by the duty cycle, which is defined as the connection period (in cell times) minus the guard time (one or two cells, for example). In a 16 slot access time with 2 slot guard time, the duty cycle is 14. The cells may belong to many outlet  
10 modules.

Figure 8 shows the operation of the middle CMs 60 in an 8x8 switch ( $N=8$ ), each row representing a middle CM which is logically divided into 8 sections, each of which corresponds to an outlet module. A section is as wide (again only logically) as the duty cycle (14 cells, for  
15 example). During an access time, inlet module 6 in Figure 8 is storing cells, in the top CM, destined to outlet modules 1, 2, 4, and 7. Each inlet module may write in different sections during the access time, after which the rotator moves to the next position. However, the accessing outlet module can only read whatever is found in its dedicated (logical)  
20 section. Thus outlet module 6 reads only cells stored in section 6 of each row as the rotator moves around.

The composition of the burst is determined through a simple matching process, as depicted in Figure 9. Each inlet module keeps an array 70 of the number of waiting cells per destination and each middle  
25 module keeps an array 72 of the number of free slots per destination. The two arrays are matched in a cyclic order. As designated by 74, the number of cells accepted is the lesser of the number of waiting cells and the number of free slots for each destination inspected, the total being limited by the duty cycle.

Figure 10 shows the simultaneous transfer of bursts from  $N$  inlet  
30 modules to  $N$  middle modules ( $P=N$ ). In the Figure, inlet modules 0, 1, ...,  $N-2$ ,  $N-1$  are accessing middle modules 1, 2, ...,  $N-1$ , 0. The logical order of rotation needs not follow the physical port order, i.e. the order can be preset in any way at the rotators. Cell order is preserved since  
35 the middle buffers are visited sequentially by both the inlet multiplexers and the outlet demultiplexers. The maximum number of cells to be stored in any middle CM is  $D$  times  $N$ , where  $D$  is the

number of cells per duty cycle. The number of cells actually stored in the middle memory varies according to traffic load composition and the cell arrival pattern.

A direct method for performing the matching process of Figures 9 and 10 is to let each inlet module send to the middle module, which it will access during the subsequent access time, an array of  $N$  words of  $d$ -bit each, e.g., an array designated by  $80$ , where  $d = \lceil \log_2(D) \rceil$ ,  $D$  being the duty cycle and  $\lceil . \rceil$  denotes rounding-up to the nearest integer. The maximum number of cells that can be transferred to any destination equals the number of cells in the duty cycle. Thus, with  $N=256$  ports and  $D=14$  (i.e.,  $d=4$ ), the number of bits transferred per access time is 1024 (approximately 2.4 cells). Each middle module then responds with a grant message 82 indicating the selected destinations, and the permissible number of cells for each. The maximum number of bits in the grant message is  $D(v + d)$  where  $v = \lceil \log_2(N) \rceil$ . (There are at most  $D$  selected outlet modules per grant;  $v$  bits store the outlet module number and  $d$  bits store the number of cells per selected outlet module.) In the above example ( $N=256$ ,  $D=14$ ), the maximum message length is 168 bits. The ratio,  $\theta$ , of the grant message overhead to the switch capacity is:

$$\theta = \frac{\lceil \log_2(N) \rceil + \lceil \log_2(D) \rceil}{B},$$

where  $B$  is the number of bits per ATM cell ( $B=424$ ). With  $N=256$  and  $D=14$ ,  $\theta$  is 0.028.

It is noted that the control-data transfer can be done in the reverse order; the middle modules may send their state information to the inlet module which performs the matching process.

This overhead (the volume of control data) can be reduced significantly by another embodiment described below. In this embodiment, the inlet modules send their inlet buffer states to middle modules every several access times and lets each middle module pass the inlet data, modified by the matching outcome, to the following unit. In other words, when middle module  $Y$  receives the state array from inlet module  $X$ ,  $Y$  performs the matching process, reduces the inlet-state array according to the outcome of the matching process, and passes the reduced array to the next middle module  $Y+1$  (modulo  $N$ ), which will be accessed by the same inlet module  $X$  during the

subsequent access time. The matching process must be implemented within the access time  $\Delta$ . It is also noted that the ratio  $\{(\Delta/D)-1\}$  is the expansion needed to realize an internally non-blocking switch (e.g., 16/14).

5 Figure 11 illustrates such a mechanism, using a ring configuration. An inlet module 90 sends its buffer state information to middle module 92, e.g.,  $CM_0$ , to which it will access for data transfer at the next access time. After having performed a matching process,  $CM_0$  updates the buffer state of the inlet module 90 and sends the  
10 information to the following module e.g.,  $CM_1$ , which performs the matching process with the updated buffer state of the inlet module 90 when the inlet module 90 accesses middle module  $CM_1$  for data transfer. The further updated buffer state information is sent to the next middle module and so on for e.g., four access times (as shown in  
15 the example of Figure 11), at which time inlet module 94 sends its buffer state information to middle module e.g.,  $CM_x$ .

This mechanism is tolerant to long propagation delays from inlet to middle modules, otherwise propagation delays exceeding the access time may complicate the control function due to the inter-  
20 dependence of successive matching processes.

Sending the inlet state data every  $L$  access times,  $L > 1$ , would reduce the corresponding control data volume by a factor  $\eta$ :

$$\eta = \frac{\lceil \log_2(L \cdot D) \rceil}{L \lceil \log_2(D) \rceil}.$$

The performance of the switch is quite insensitive to this artificial  
25 delay, and one may choose to send the inlet update every 16 or 32 access times. With  $L=32$  and  $D=14$ , the factor  $\eta$  is 0.0703. Note that  $\eta=1$  when  $L=1$ . The ratio,  $\epsilon$ , of the control data overhead from the inlet modules to the middle modules to the switch capacity is:

$$\epsilon = \frac{N \lceil \log_2(L \cdot D) \rceil}{B L D},$$

30 where  $B$  is number of bits per ATM cell ( $B=424$ ). With  $L=32$ ,  $D=14$ , and  $N=256$ ,  $\epsilon$  is only 0.012.

The variable delay encountered in traversing the inlet and middle CMs is negligible for all traffic streams. Hence, priority  
35 classification is not necessary in either the inlet stage or the middle stage. It is noted, however, that priority service may be needed at

egress, i.e., in the outlet stage. The egress performance is similar to that of a single-stage CM switch and is not discussed here.

The relevant performance indices here are the grade-of-service (GOS), determined mainly by the call-admission blocking and the  
5 quality-of-service (QOS), which is determined by the cell loss and/or cell transfer delay. The cell delay and cell loss contribution of the switching network of the present invention is at least an order of magnitude smaller than the contribution of the egress stage. Thus the overall performance is comparable to that of the (ideal) single stage  
10 switch, under similar traffic conditions.

The performance issues at the call and cell levels will be discussed below. Call-level blocking applies mainly to CBR and VBR traffic streams where the admission of a new arrival depends on its declared traffic descriptors and, hence, a calculated "effective bit rate  
15 (EBR)". The EBR value is determined by the cell-level performance (cell loss and/or cell delay variation). The multiplexing of the lower rate traffic at the inlet modules into a high speed stream, and the subsequent distribution among the middle CMs, have very little effect on the overall cell delay which is determined mainly by the outlet  
20 occupancy. The internal variable delay is negligible since each inlet module is free to transfer cells through any middle module for any virtual circuit. In addition, as seen in Figure 16 which will be described later, the modest cell storage requirement at the inlet modules eliminates cell loss as a major concern. Thus, the EBR calculation can  
25 be based on standard methods applicable to single-stage output buffered or common memory switches. Like the output buffered or common memory switches, the switches of the present invention, with zero guard time, can be treated as non-blocking at the call level. In other words, the admission, or otherwise, of an arrival is determined only by  
30 the state of the designated outlet.

With a non-zero guard time, the link capacity is somewhat reduced. It is customary, however, to allow some internal expansion where the inner links are of a slightly higher speed than the outer links. The expansion is provided to facilitate internal flow control and  
35 it does not affect the traffic performance of single-stage switches. The expansion has the added benefit of offsetting the guard time overhead.

By definition, a switch is considered non-blocking if the blocking of an incoming request is determined solely by the designated outgoing link. Because of link blocking, the call-level occupancy of an outlet port, i.e., the sum of the EBRs of the calls in progress divided by the port bit rate, fluctuates around its mean value below unity. The outer links would occasionally be in the state of full call level occupancy. To realize an acceptable call blocking (0.01 for example), the outer links may be engineered for a mean call level occupancy of 0.8 or so, depending on the traffic composition. The mean cell level occupancy is lower than the mean call level occupancy since the EBR for a VBR connection is always higher than the mean bit rate of the connection. The simulation results which will be discussed later are based on a pessimistic mean cell level occupancy of 0.80, and the internal expansion is assumed to be zero. With a typical expansion of 0.1 or so, the delay variation would be appreciably smaller.

Cell level performance is normally expressed in terms of the cell loss probability and the cell delay variation. The cross office round trip delay, traditionally specified for circuit switches to be less than one millisecond or so, is still applicable to ATM switches.

The cross office round trip delay for the proposed switch is a constant which is equal to the rotator cycle duration. In a 256 port switch, with a port speed of 10 Gb/s, the rotator cycle is about 175  $\mu$ sec with a 16 slot access time, or 88  $\mu$ sec with an access time of 8 slots. The round trip delay is the sum of the delay from inlet port  $x$  to outlet port  $y$  plus the delay from inlet port  $y$  to outlet port  $x$ . The two components are not equal, and each varies from one access time  $\Delta$  to  $(N-1)\Delta$ . A long  $x$ - $y$  delay corresponds to a short  $y$ - $x$  delay, and the sum is constant.

The cell-delay variation is the more critical performance index since it determines the size of the smoothing buffers used for CBR connections. The switch of the invention (with 5 to 10 Gb/s rotator links) yields a delay dispersion, at the 10<sup>th</sup> quantile, well below the commonly accepted bound of 250  $\mu$ sec at the chosen reference load.

According to yet a further embodiment, the folded arrangement depicted in Figure 12 may be used in the configurations of Figures 2, 3, 6 and 7. Each CM module serves as a combined multiplexer-demultiplexer-switch as in Figure 13. An  $n \times n$  CM module is operated



as an  $\frac{n}{2} : 1$  multiplexer, a  $1 : \frac{n}{2}$  demultiplexer, and an  $\frac{n}{2} : \frac{n}{2}$  switch (for an even number of external ports  $n$ ). The effective internal expansion ratio is increased due to the intra-switching facility.

5 Since the multiplexing function requires very little storage capacity, the common memory capacity of each combined inlet-outlet module is used mainly for egress queuing. A major advantage of this configuration is that only the inter-module traffic would have to traverse the middle stage. This results in reducing the rate of cell transfer across the middle stage and hence decreasing the contention  
10 delay.

The set of virtual links connecting the outer CMs and the middle CMs is realized as a simple rotator. According to further embodiments of the invention, large rotators may be constructed using smaller size rotator units of size  $k \times k$  each ( $k > 1$ ) by cascading banks of  
15 small units operating at different speeds, that is to say, an  $k^h \times k^h$  rotator  $k=1,2,\dots$ , can be built by using  $h$  rows of  $k^{h-1}$  smaller rotators of size  $k \times k$  each. The innermost units must switch ports every  $\Delta$  slots, where  $\Delta$  is the desirable access-time. The units of the second bank must switch ports every  $k\Delta$  slots. Figure 14 shows a two-stage configuration which  
20 extends the capacity to  $k^2 \times k^2$ . A third bank, whose units switch ports every  $k^2\Delta$ , extends the capacity to  $k^3 \times k^3$  as shown in Figure 15. For example, a  $256 \times 256$  rotator requires 32 rotators of size  $16 \times 16$  arranged in two rows of 16 units each (here  $k=16$  and  $h=2$ ). In the configuration of Figure 15, with  $k=16$ , a  $4096 \times 4096$  rotator can be constructed with 768  
25 units (three rows  $h=3$  of 256 units) of  $16 \times 16$  rotators. It is interesting to note that a non-blocking space switch of the same size would require 65536 units of  $16 \times 16$  space switches arranged as a square. Unlike the space switch, the rotator's operation is cyclic and traffic independent.

### 30 Simulation Results

A simulator for the proposed architecture was developed and used to study the performance of switches of different sizes ranging from  $N=8$  to  $N=256$ , with both the folded architecture (with intra-switching in the outer modules) and the unfolded architecture (with  
35 no intra switching). The number of cells processed in each case is about

$2.5 \times 10^8$ . The inlet module buffer occupancy and the variable cell delay are shown for a port mean cell-occupancy of 0.80.

The traffic arriving at an inlet module port (of OC12 rate, for example) is a multiplex of traffic streams generated by several sources.

5 The traffic generated by each source is assumed to be very bursty, with a large ratio of peak rate to mean rate. Using the ON-OFF model with geometrically distributed "ON" and "OFF" periods, the multiplexed traffic at the inlet module port is assumed to have a mean burst length of 20 (implicitly, the individual sources would have much larger burst  
10 lengths at their own peak rates). The composite traffic, at the rotator port rate of 10 Gb/s or so, is much less bursty. As shown in the simulation results below, the dispersion of the delay of cell transfer to the output stage is quite small. The egress process at the output stage (demultiplexing to slower ports) contributes most of the delay and is  
15 affected by burstiness in the same way as a single-stage CM switch.

The simulation results were derived for spatially-balanced traffic. Results obtained for several cases with high spatial imbalance (large variance of traffic intensity for different inlet-outlet pairs) show negligible sensitivity to the spatial traffic distribution as long as the  
20 overall load for each outlet port remains unchanged.

Figure 16 shows the inlet buffer distribution for the case of access time  $\Delta$  of 16 cell intervals (slots) with a guard time  $\gamma$  of 2 cell intervals. For a cell loss of the order of  $10^{-7}$ , a buffer size of less than 35 cells suffices.

25 Figure 17 shows the distribution of the variable cell delay (cell delay complementary function) in the folded architecture. The access time  $\Delta$  is kept constant at 16 cell intervals for the different switch sizes. The guard time is 2 and the cell interval is 0.08  $\mu$ seconds. The delay increases with the increase of the number of ports due to the increase  
30 in the rotator cycle.

Figure 18 shows the delay performance (cell delay complementary function) in the folded architecture when the access time  $\Delta$  is reduced to 8 slots and the guard time is reduced to one slot. The cell interval is 0.08  $\mu$ seconds. The idle (guard) time remains  
35 proportionately the same as in the case of Figure 17, however, the delay performance improves due to the reduced access time.

Figure 19 shows the delay performance (cell delay complementary function) in the case of the unfolded architecture (no intra-module switching). The access time  $\Delta$  is 16 cells, the guard time is 2 cells and the cell interval is 0.04  $\mu$ second.

## WHAT IS CLAIMED IS:

1. In a rotating access high capacity ATM switching system for switching data among N inlet modules and M outlet modules in each successive access time in a burst of a predetermined number of cells, M and N being positive integers, the invention being characterized in that:
  - said N inlet modules having buffers, each buffer dedicated to each of said outlet modules, for storing cells according to destination outlet modules of said cells in respective buffers;
  - P common memories, P being a positive integer, each common memory having M memory sections, each of which is able to hold at least said predetermined number of cells and is dedicated to each outlet module;
  - an inlet rotator for cyclically connecting in each access time said N inlet modules and P common memories so that respective cells are transferred from said N inlet modules and stored in respective sections according to the destination outlet module of each cell; and
  - an outlet rotator for cyclically connecting in each access time said P common memories and M outlet modules so that respective outlet modules are connected to respective memory sections for reading out cells contained therein.
2. The high capacity ATM switching system according to claim 1 wherein the inlet module comprises a buffer which is logically partitioned to buffer sections corresponding to said M output modules.
3. The high capacity ATM switching system according to claim 2 wherein said inlet and outlet rotators are  $k \times k$  rotators making k simultaneous connections, k being an integer larger than 1.
4. The high capacity ATM switching system according to claim 3 further comprising a matching mechanism for matching cells stored in the buffers of the inlet modules and free slots in the sections of the common memories according to the destinations of the predetermined number of cells in a burst.

5. The high capacity ATM switching system according to claim 4 wherein the inlet modules send inlet control data, to the matching mechanism, concerning the number of cells stored in their buffers according to the destinations of the predetermined number of cells; the common memories send memory state data, to the matching mechanism, concerning the number of free slots available in the sections; and the matching mechanism sends grant signals to the inlet modules for the number of cells to be transferred according to the destinations of the predetermined number of cells.

6. The high capacity ATM switching system according to claim 4 further comprising:

each inlet module sending inlet control data to one of the common memories in every several access times;

a ring controller connecting said P common memories in a ring configuration;

each common memory having a matching mechanism for sending to each inlet module the grant signals during each access time and updating the inlet control data as a result of the grant signals; and

each common memory sending the updated inlet control data to the following common memory in the ring.

7. The high capacity ATM switching system according to claim 5 wherein  $M=N=P$ .

8. The high capacity ATM switching system according to claim 6 wherein  $M=N=P$ .

9. The high capacity ATM switching system according to claim 7 wherein each of the input and output rotators comprises h tandemly connected sets of  $(k \times k)$  rotators where h is a positive integer and  $k = \left\lceil M^{\frac{1}{h}} \right\rceil$  in that the number of said  $(k \times k)$  rotators in each set is k and one set operates k times faster than the other set.

10. The high capacity ATM switching system according to claim 8 wherein each of the input and output rotators comprises h tandemly

connected sets of  $(k \times k)$  rotators where  $h$  is a positive integer and  $k = \left\lceil M^{\frac{1}{h}} \right\rceil$  in that the number of said  $(k \times k)$  rotators in each set is  $k$  and one set operates  $k$  times faster than the other set.

- 5            11. In a rotating access high capacity ATM switching system for switching data among  $N$  inlet modules and  $M$  outlet modules in each successive access time,  $M$  and  $N$  being positive integers, a method of switching data in a burst of a predetermined number of cells among  $N$  inlet modules and  $M$  outlet modules in each successive access time  
10    being characterized in that:  
             each of said  $N$  inlet modules storing cells in separate buffers according to the destination outlet modules of said cells;  
             cyclically connecting said  $N$  inlet modules and  $P$  common memories,  $P$  being a positive integer;  
15    transferring in each access time the burst of said predetermined number of cells from one of said  $N$  inlet modules to respective memory sections of one of said common memories according to the destination outlet modules of said cells; and  
             cyclically connecting said common memories and  $M$  outlet  
20    modules so that respective outlet modules are connected to a respective memory section for reading out cells contained therein.
12. The method of switching data in a burst of a predetermined number of cells according to claim 10, wherein the steps of cyclically  
25    connecting comprises steps of:  
             cyclically making  $k$  simultaneous connections in each access time between said inlet modules and said common memories, and between said common memories and  $M$  outlet modules,  $k$  being an integer larger than 1.  
30
13. The method of switching data in a burst of a predetermined number of cells according to claim 12 further comprising steps of:  
             matching cells stored in the buffers of the inlet modules and free  
35    slots in the sections of the common memories according to the destination of the predetermined number of cells in a burst.

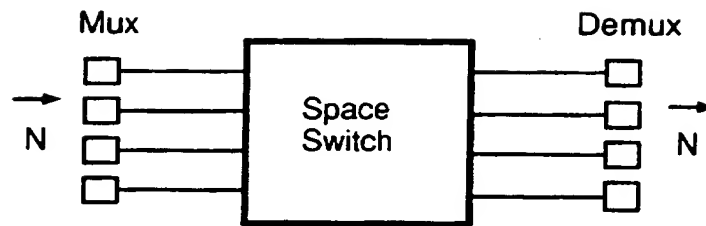
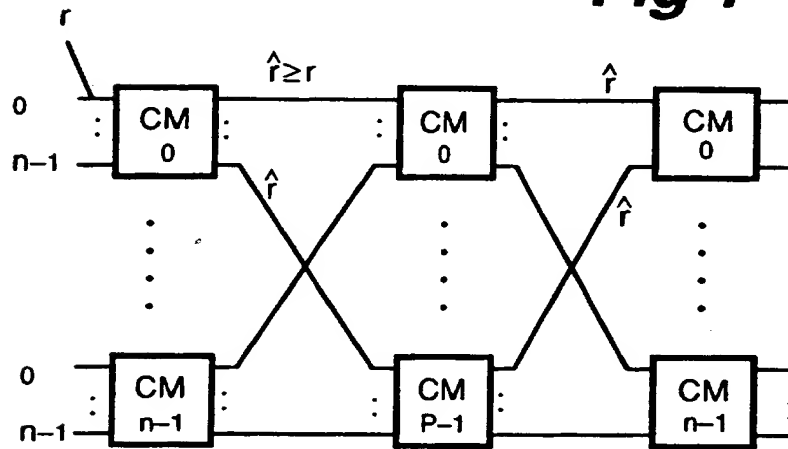
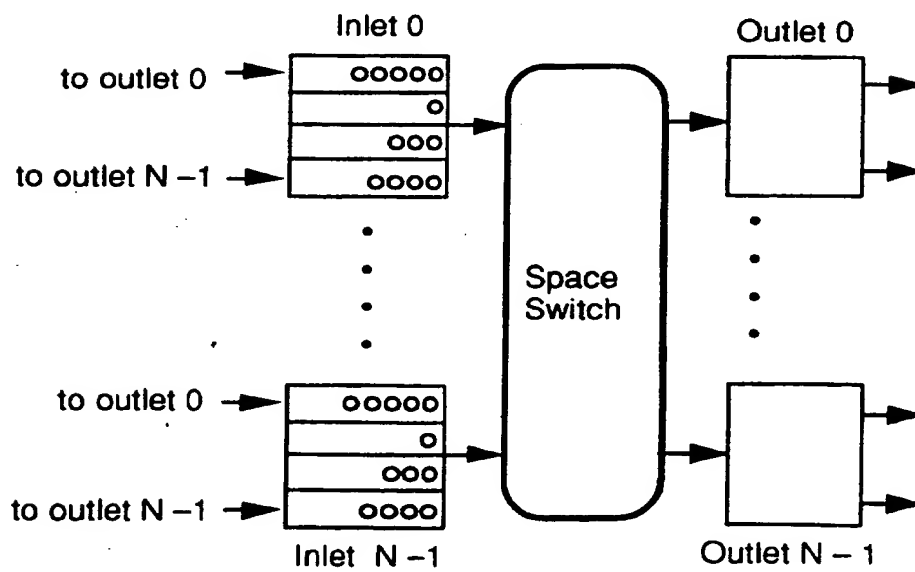
14. The method of switching data in a burst of a predetermined number of cells according to claim 13, further comprising steps of:

- 5 matching inlet control data and memory state data, the former concerning the number of cells stored in their buffers according to the destinations of the predetermined number of cells and the latter concerning the number of free slots available in the sections; and
- 10 sending grant signals to the inlet modules for a number of cells to be transferred according to the destinations of the predetermined number of cells.

15 15. The method of switching data in a burst of a predetermined number of cells according to claim 14 wherein said P common memories are connected in a ring configuration, the method further comprising steps of:

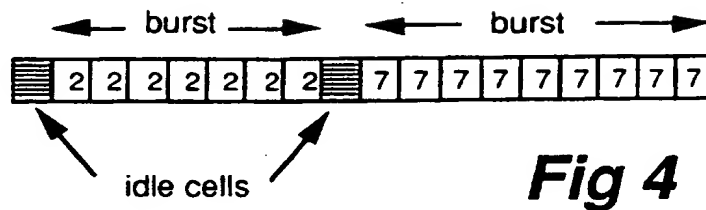
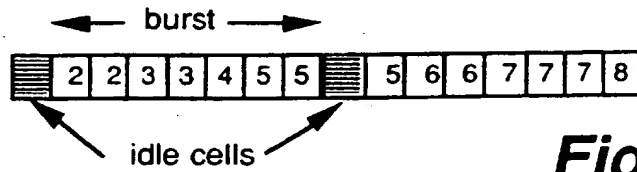
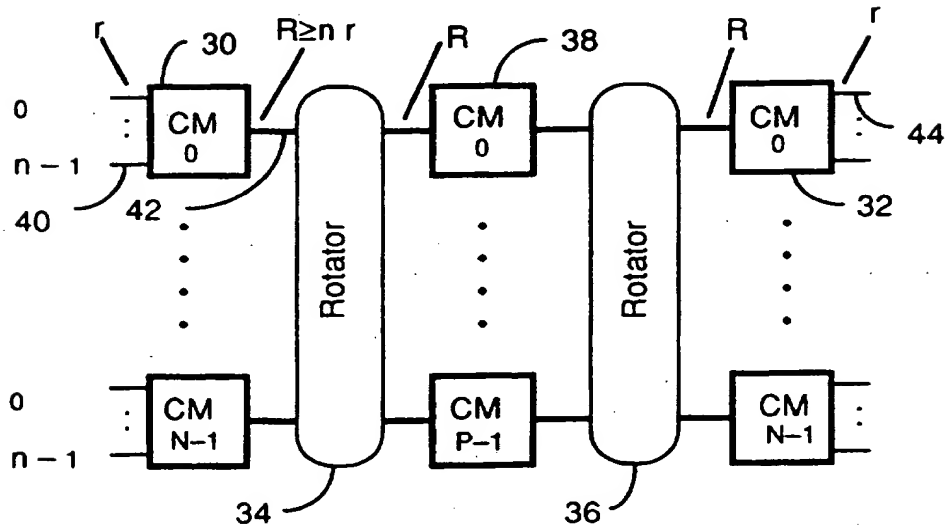
- 15 each inlet module sending inlet control data to one of the common memories in every several access times;
- each common memory matching the inlet control data and memory state data and sending to each inlet module the grant signals during each access time;
- 20 each common memory further updating the inlet control data as a result of the grant signals; and
- each common memory sending the updated inlet control data to the following common memory in the ring.

1/10

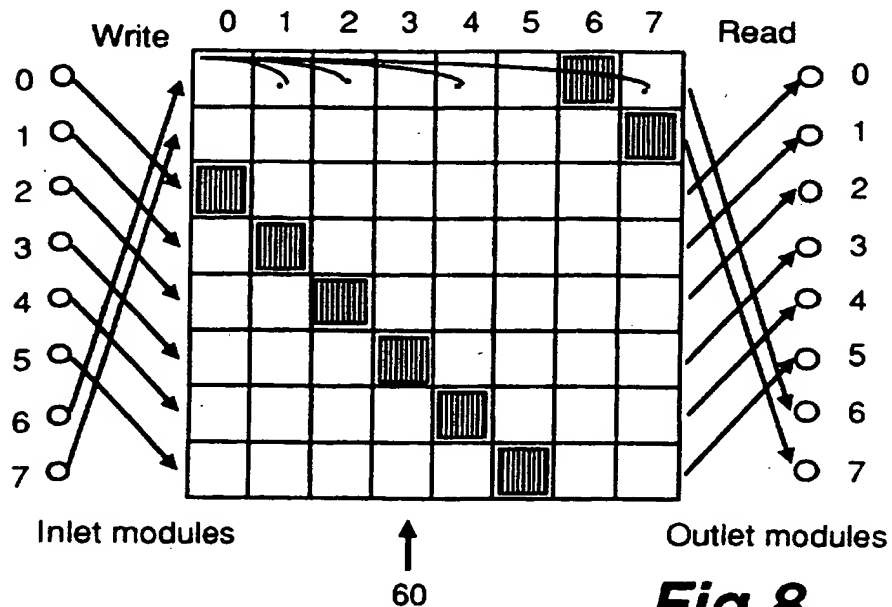
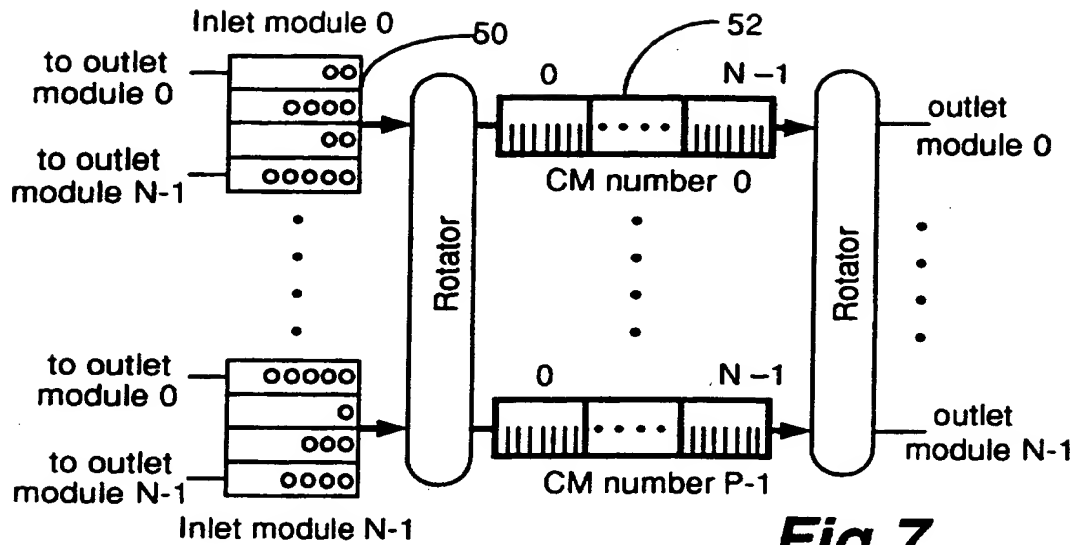
**Fig 1****Fig 2****Fig 3**



2/10

**Fig 4****Fig 5****Fig 6**

3/10



4/10

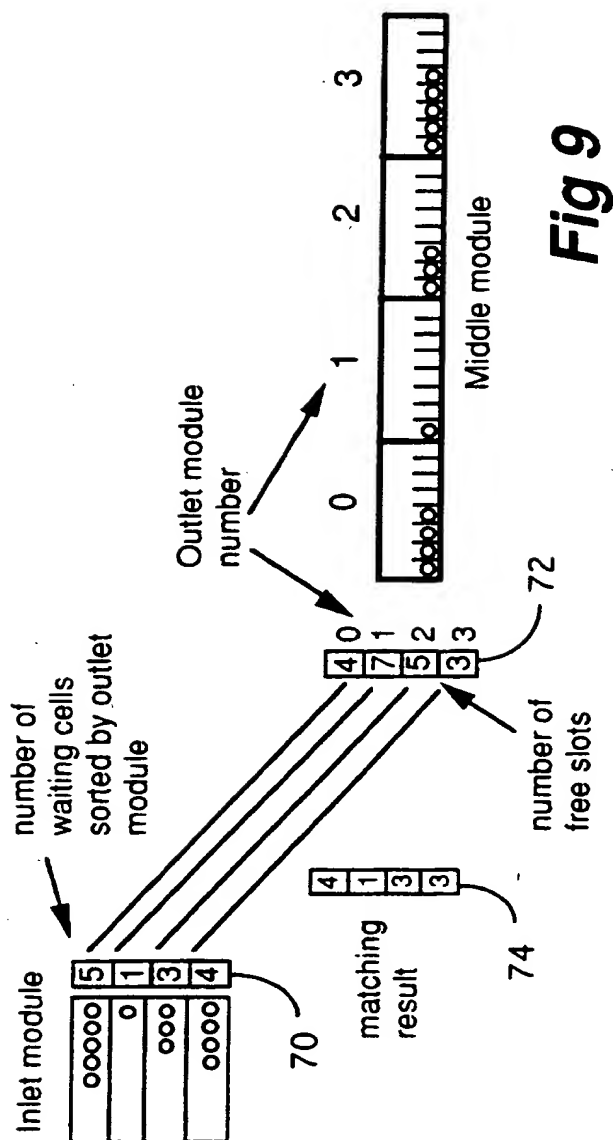


Fig 9

5/10

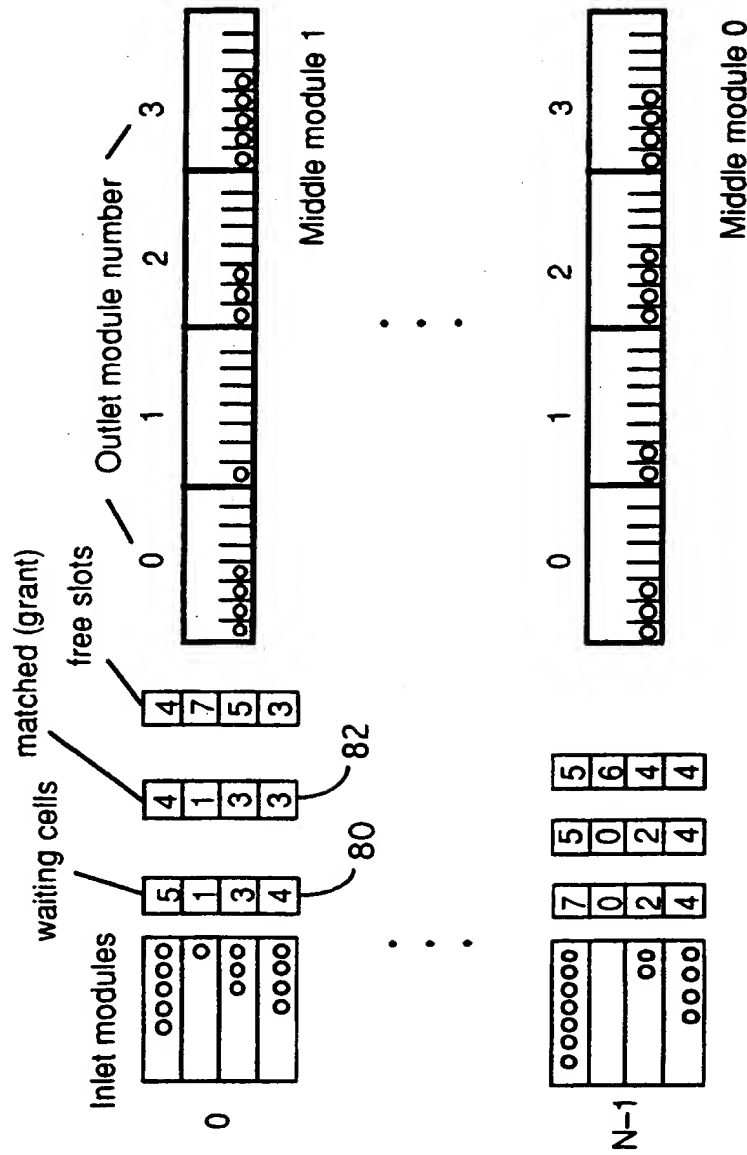
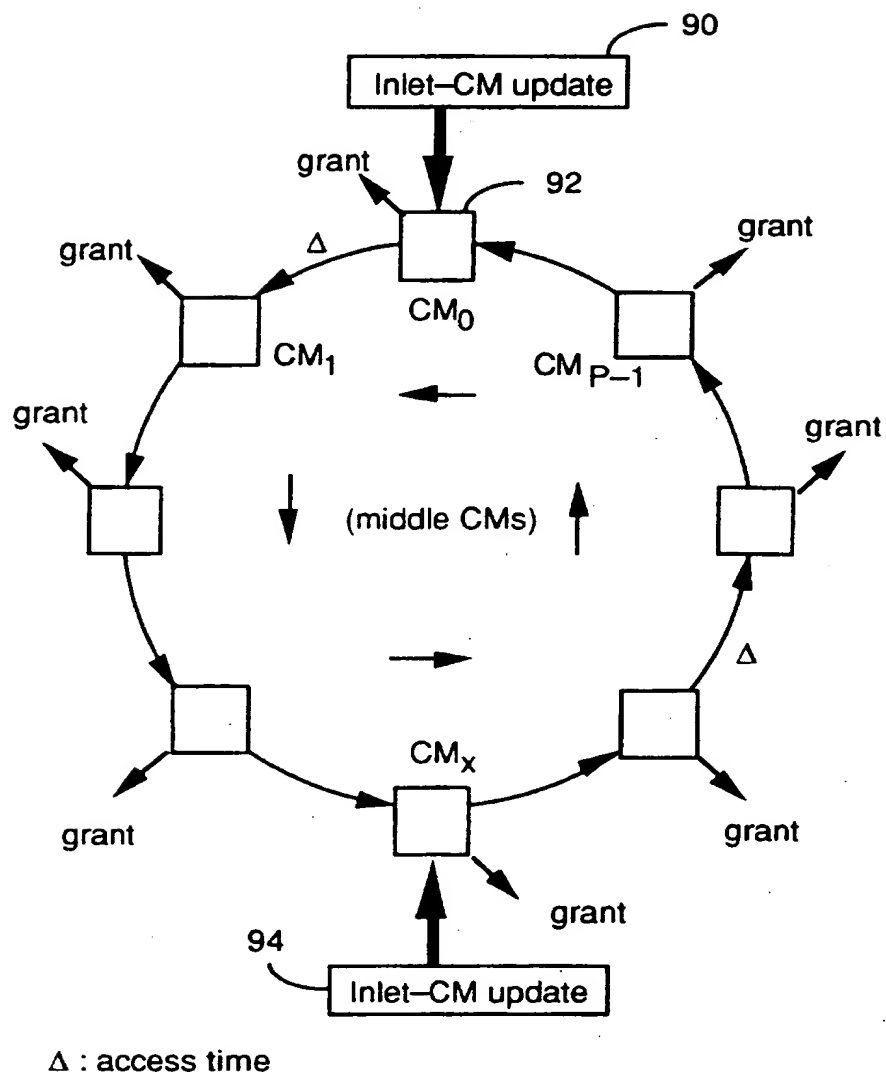
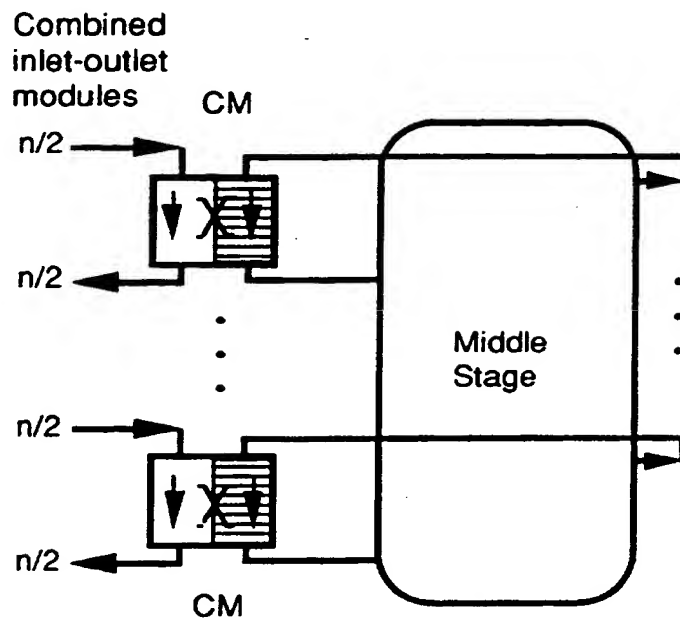
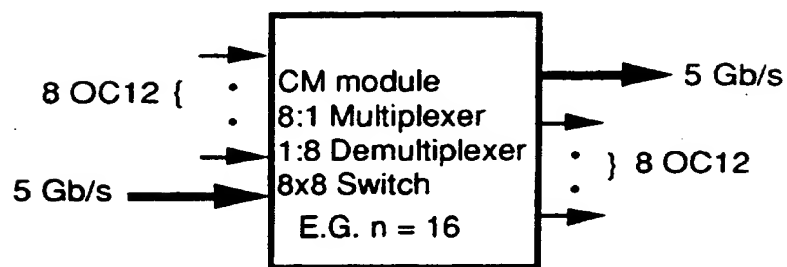


Fig 10

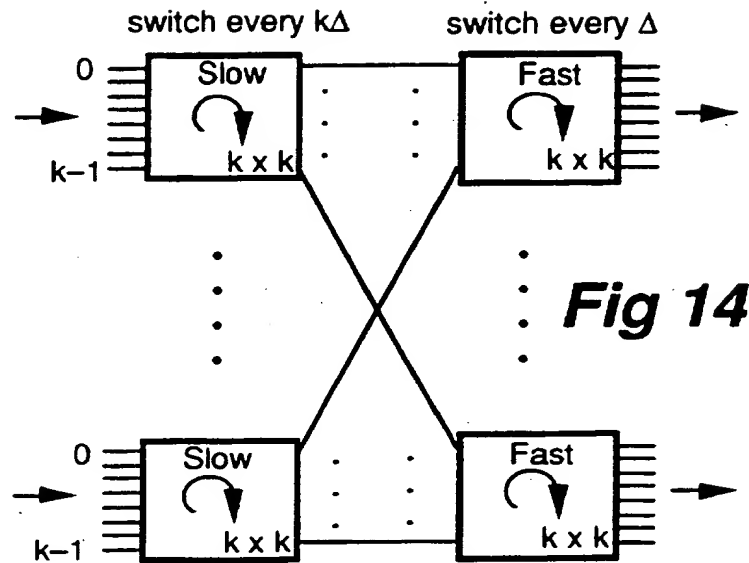
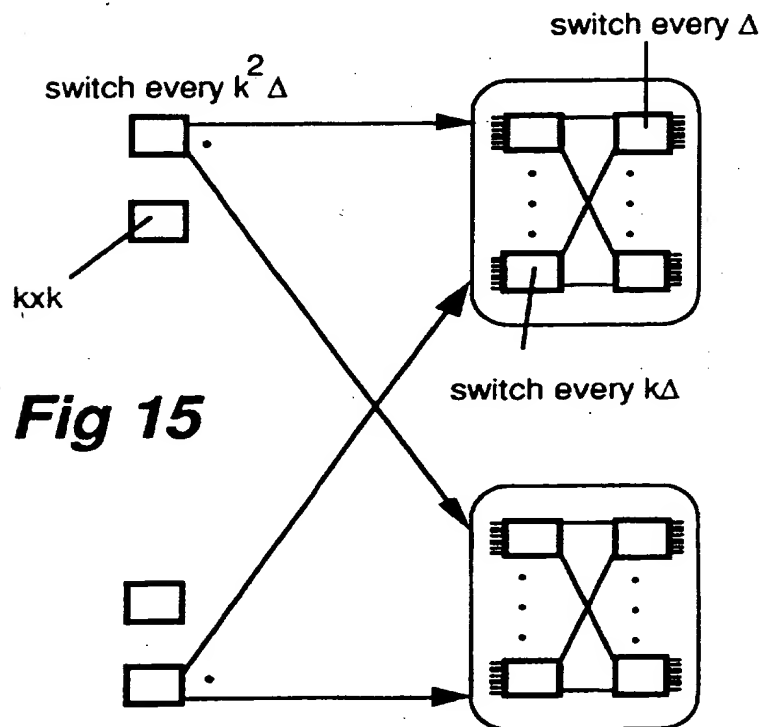
6/10

**Fig 11**

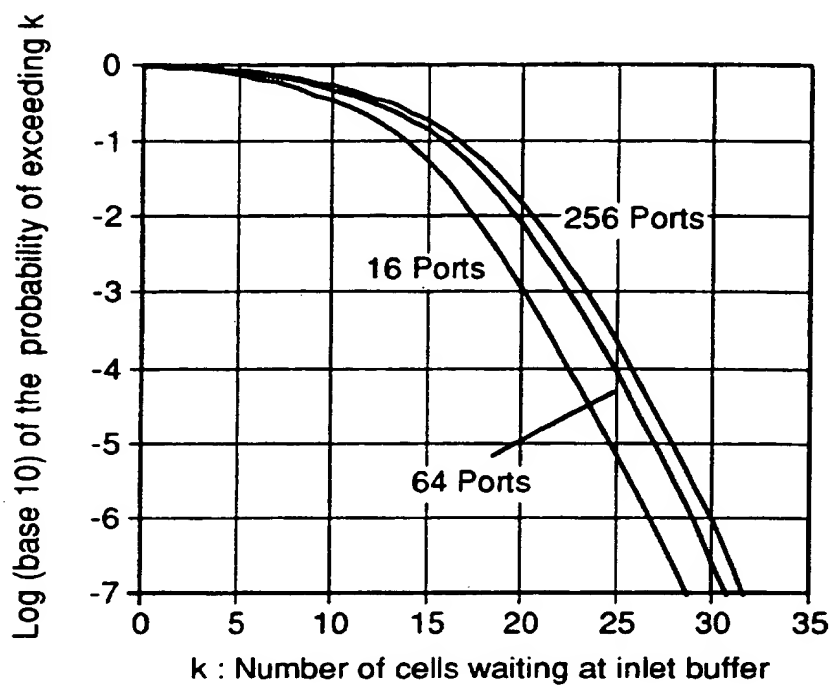
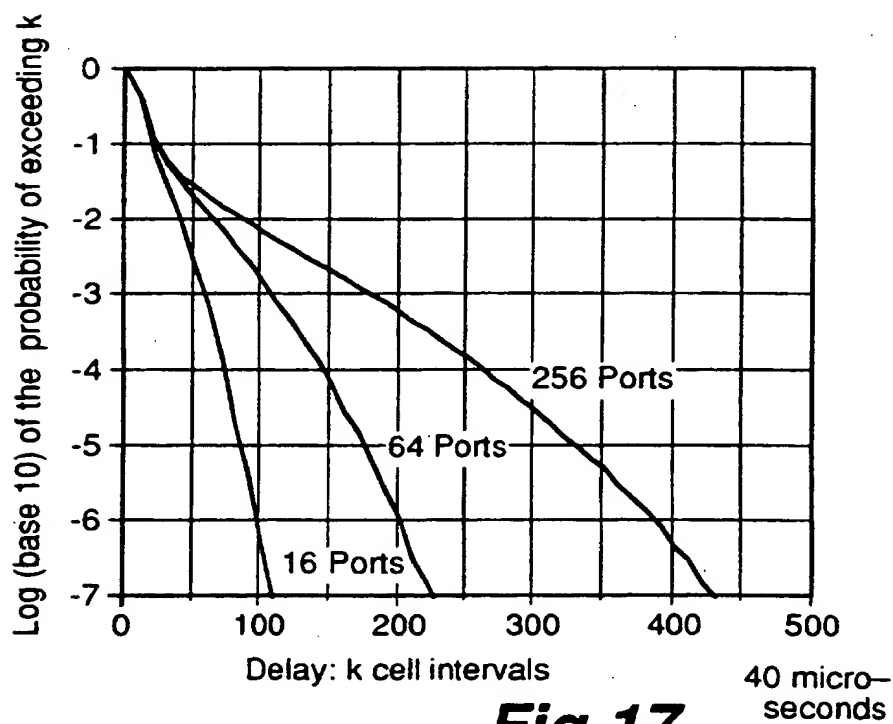
7/10

**Fig 12****Fig 13**

8/10

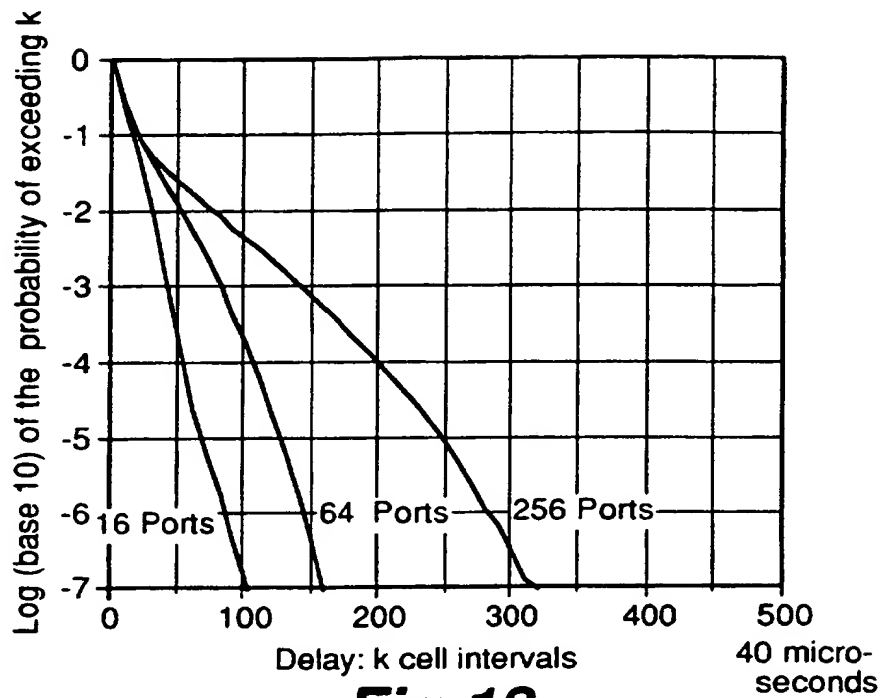
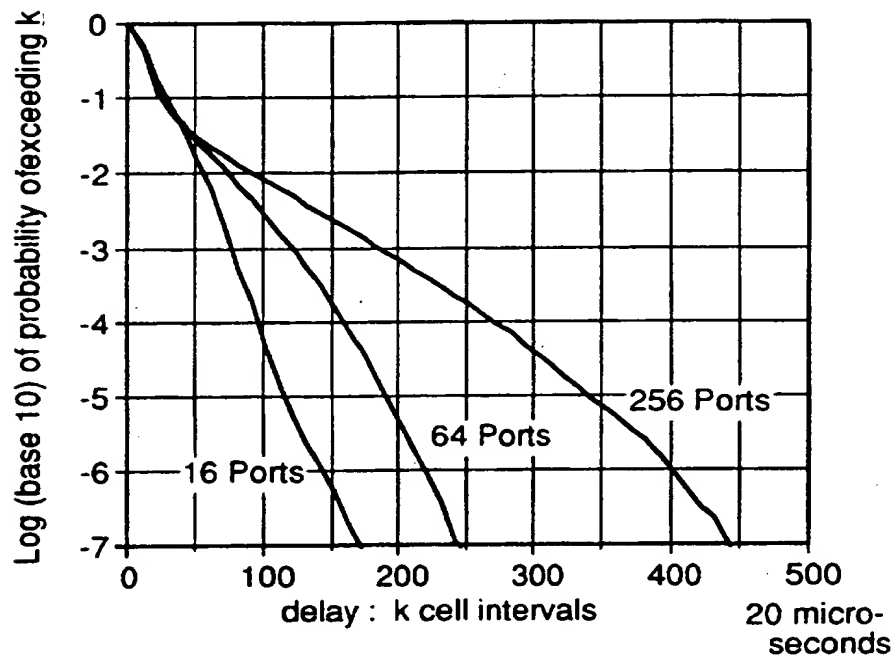
**Fig 14****Fig 15**

9/10

**Fig 16****Fig 17**



10/10

**Fig 18****Fig 19**

## INTERNATIONAL SEARCH REPORT

International Application No

PC, CA 96/00673

A. CLASSIFICATION OF SUBJECT MATTER  
IPC 6 H04L12/56

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 6 H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5 168 492 A (BESHAÏ MAGED E ET AL) 1 December 1992 cited in the application	1-3, 11, 12
A	* Figure 2 * see column 3, line 6-19 see column 4, line 10-50 see column 4, line 63-67 see column 7, line 15-20 --- -/--	4-10, 13-15

☒ Further documents are listed in the continuation of box C.☒ Patent family members are listed in annex.

\* Special categories of cited documents:

- \* 'A' document defining the general state of the art which is not considered to be of particular relevance
- \* 'E' earlier document but published on or after the international filing date
- \* 'L' document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- \* 'O' document referring to an oral disclosure, use, exhibition or other means
- \* 'P' document published prior to the international filing date but later than the priority date claimed

\* 'T' later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

\* 'X' document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

\* 'Y' document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

\* '&amp;' document member of the same patent family

Date of the actual completion of the international search

24 January 1997

Date of mailing of the international search report

04.02.97

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+ 31-70) 340-2040, Tx. 31 651 epo nl,  
Fax: (+ 31-70) 340-3016

Authorized officer

Dhondt, E

## INTERNATIONAL SEARCH REPORT

International Application No.

PC 1/CA 96/00673

## C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

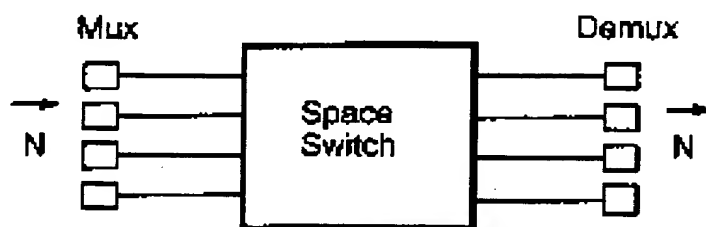
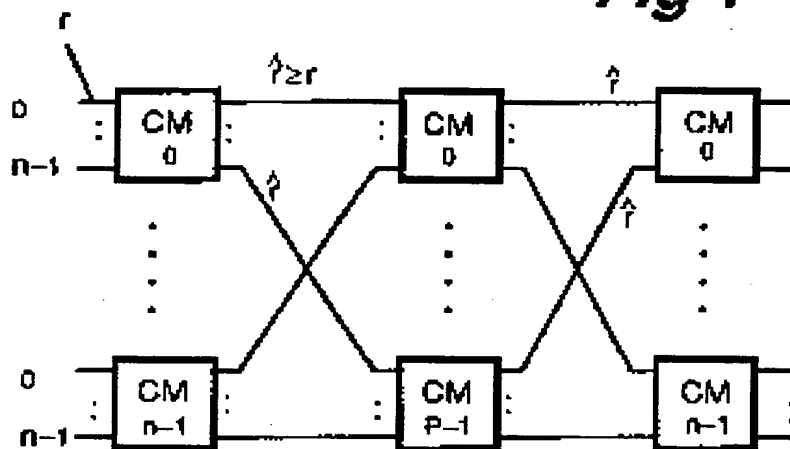
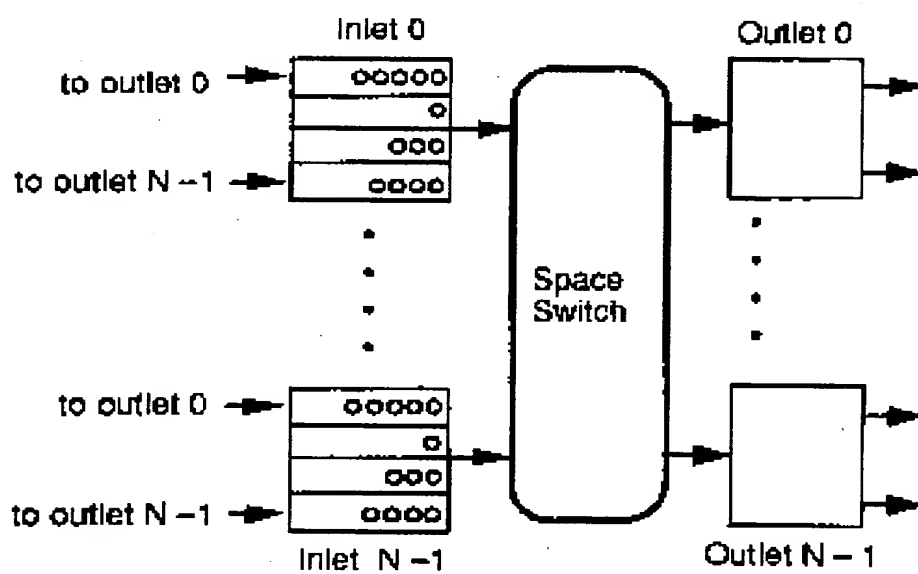
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	SERVING HUMANITY THROUGH COMMUNICATIONS. SUPERCOMM/ICC, NEW ORLEANS, MAY 1 - 5, 1994, vol. VOL. 1, no. -, 1 May 1994, INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, pages 402-409, XP000438948 FAN R ET AL: "EXPANDABLE ATOM SWITCH ARCHITECTURE (XATOM) FOR ATM LANS" see figure 5 see page 406, column 1, line 1-5 ---	1-3,11, 12
A	TECHNOLOGY AND FOUNDATIONS. INFORMATION PROCESSING '94, HAMBURG, AUG. 28 - SEPT. 2, 1994, vol. 1, 28 August 1994, PEHRSON B;SIMON I (EDITORS), pages 203-210, XP000478806 DREFENSTEDT R ET AL: "APPLICATIONS OF PRAMS IN TELECOMMUNUCATIONS" see figure 2 see paragraph 4 -----	1,11

**Information on patent family members**

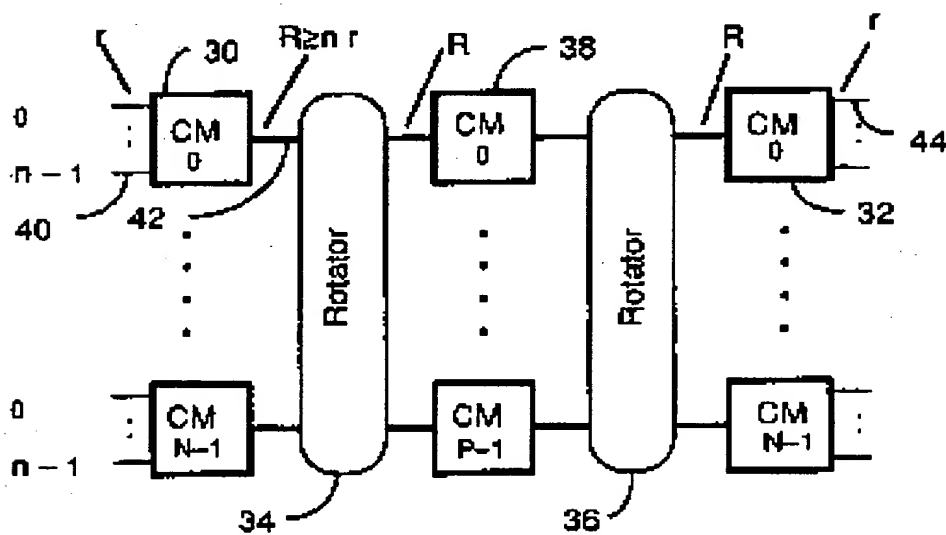
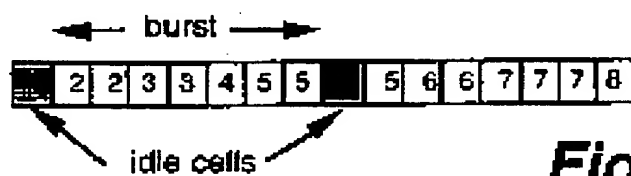
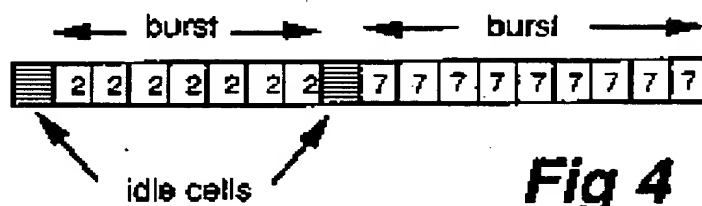
PCi/CA 96/00673

Form PCT/ISA/210 (patent family annex) (July 1992)

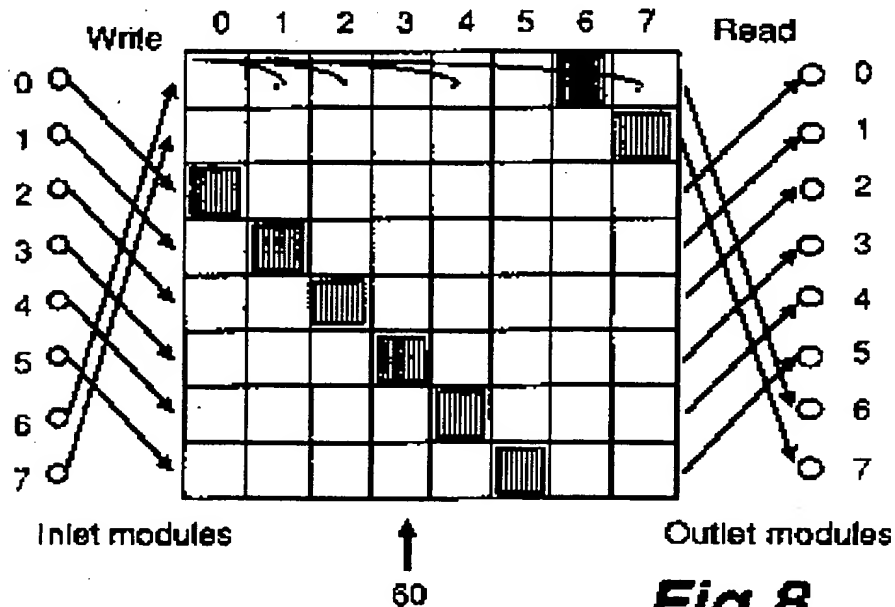
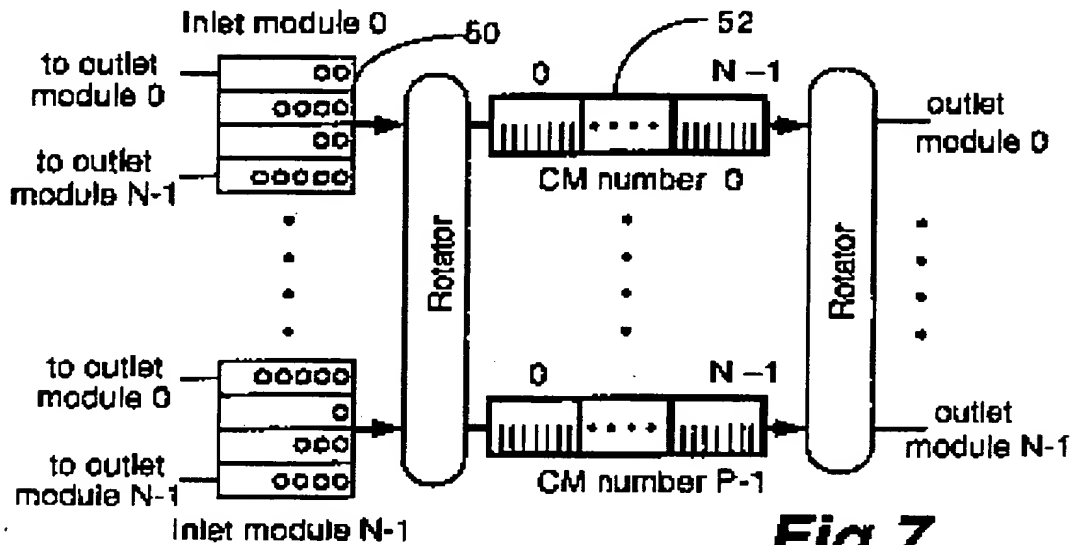
1/10

**Fig 1****Fig 2****Fig 3**

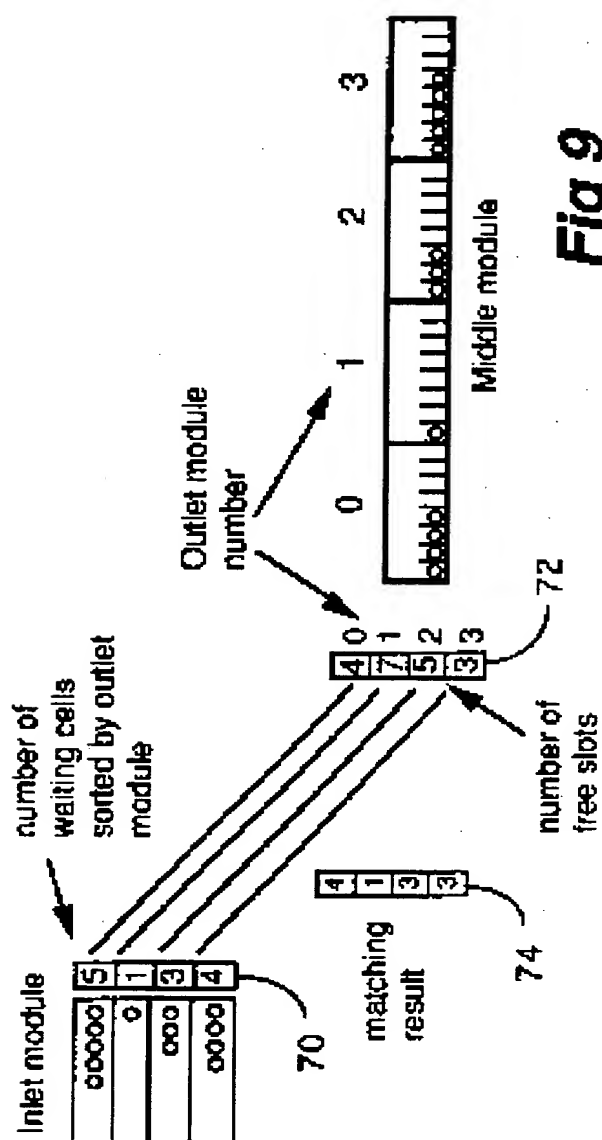
2/10



3/10

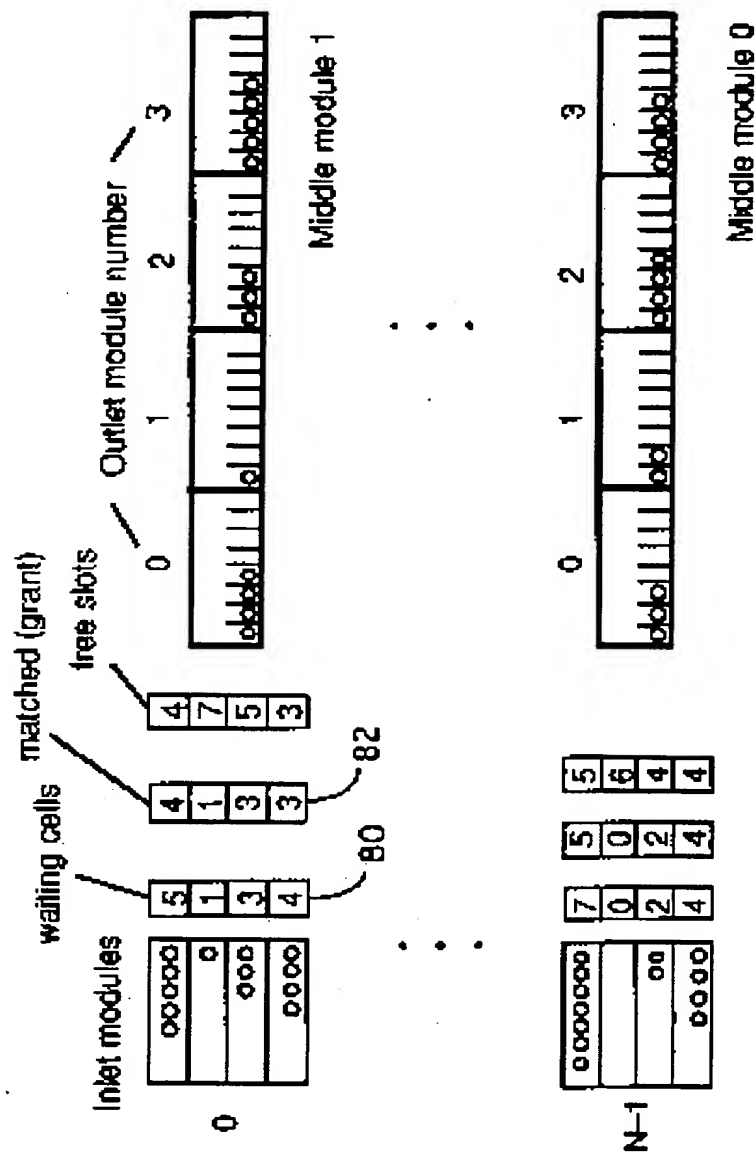


4/10

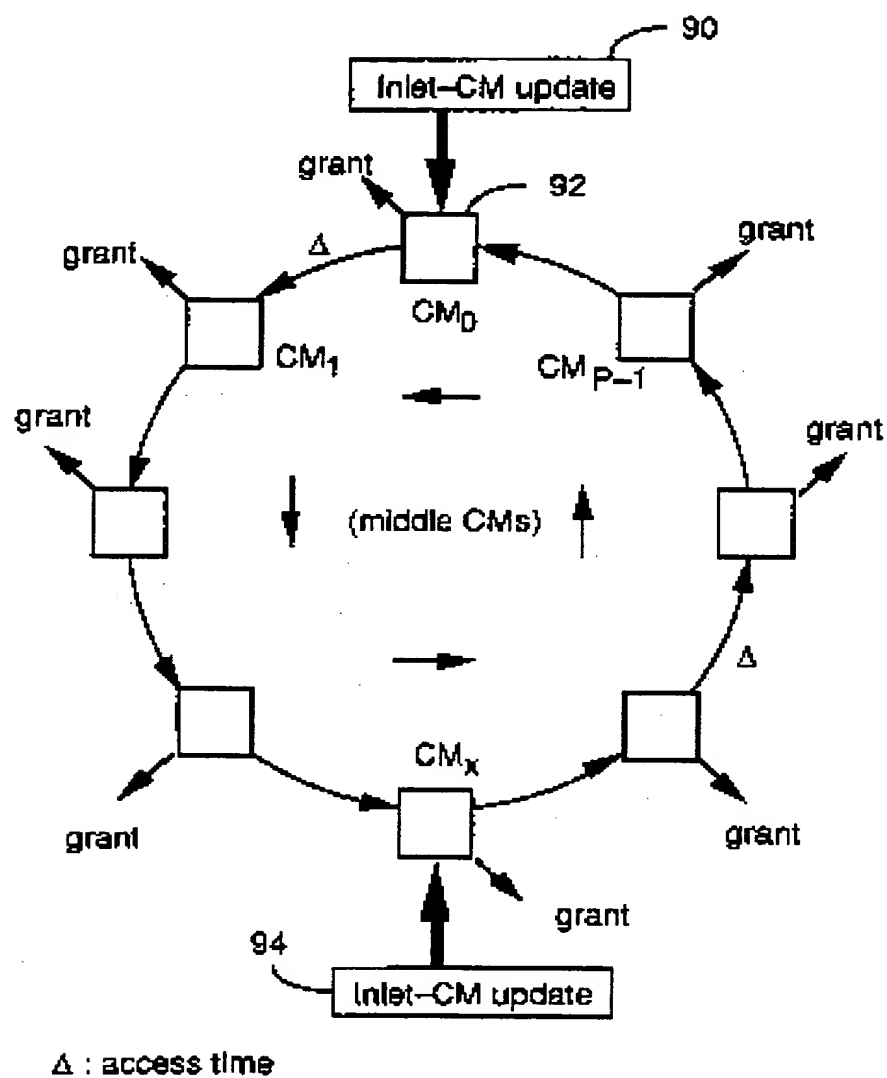
**Fig 9**



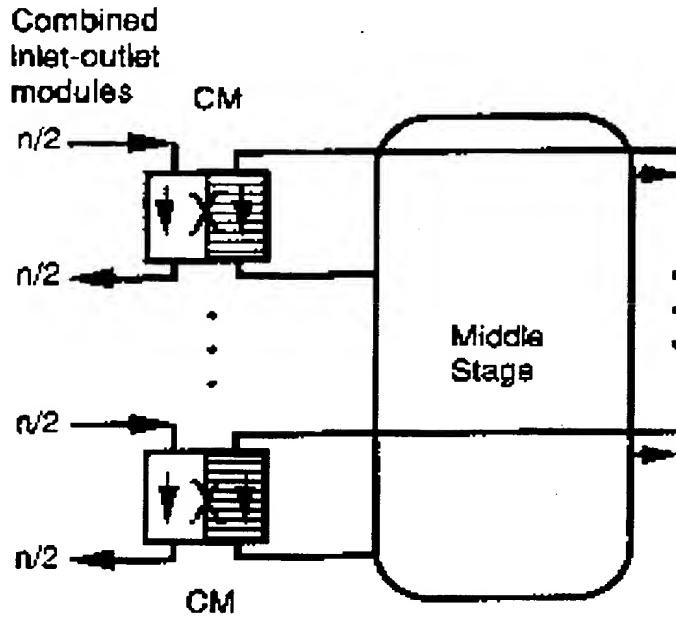
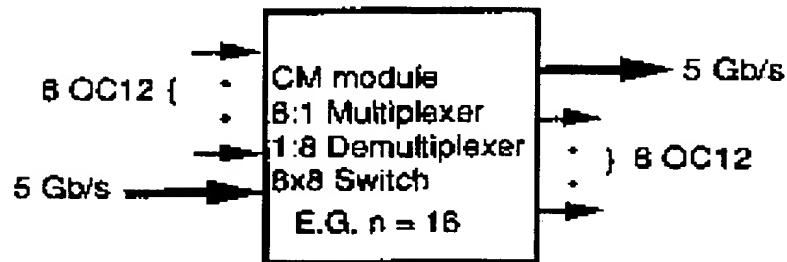
5/10

**Fig 10**

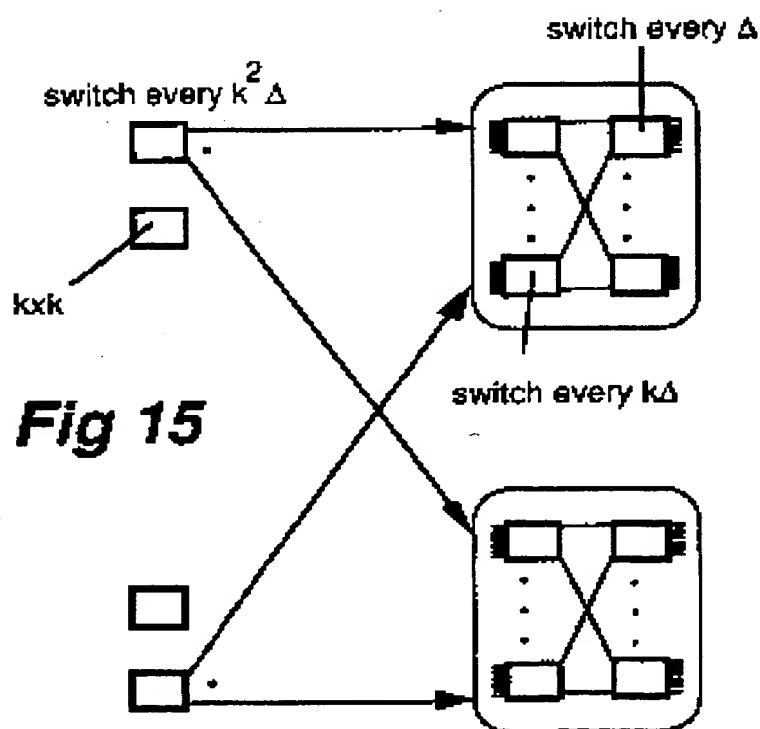
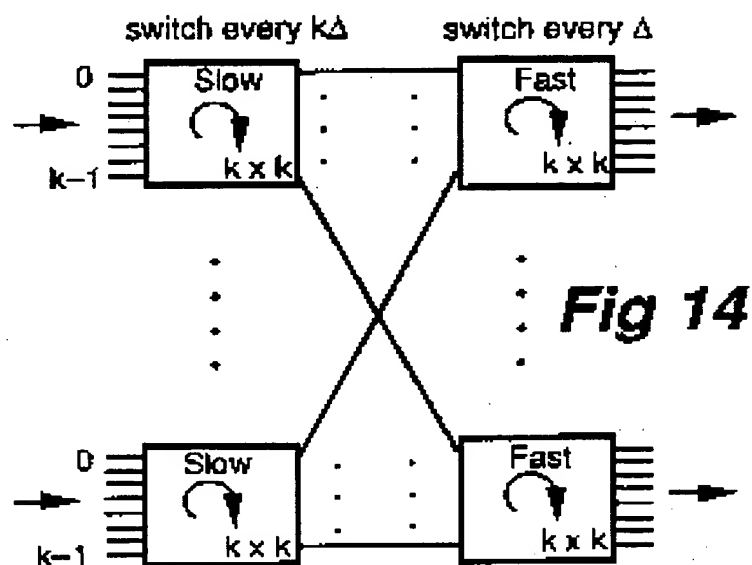
6/10

**Fig 11**

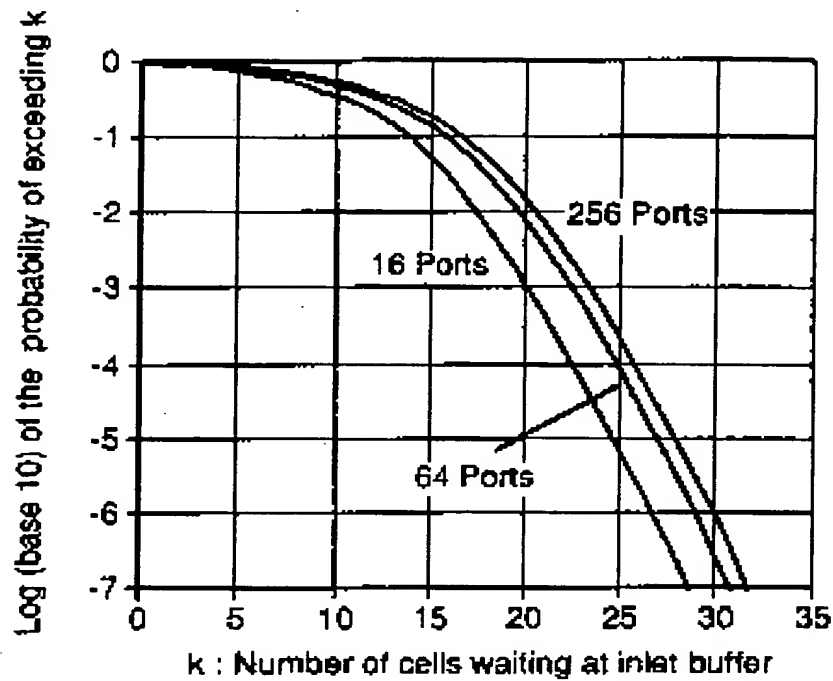
7/10

**Fig 12****Fig 13**

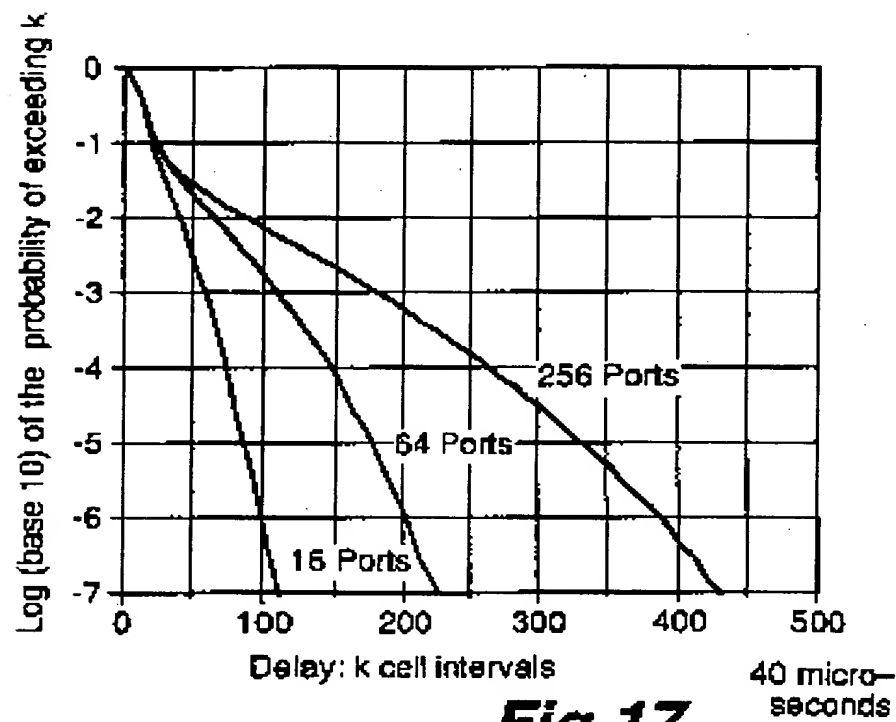
8/10



9/10

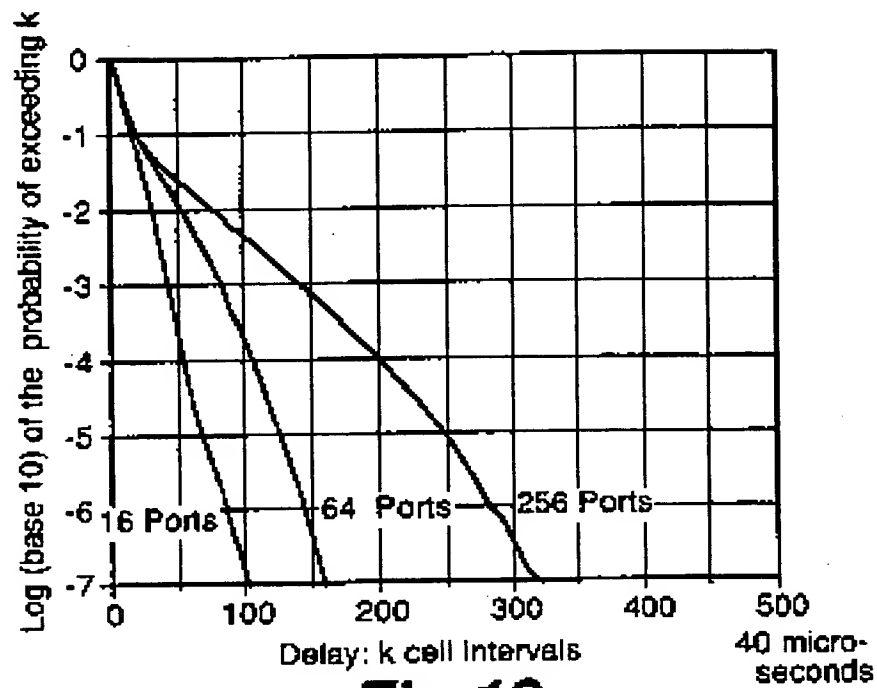
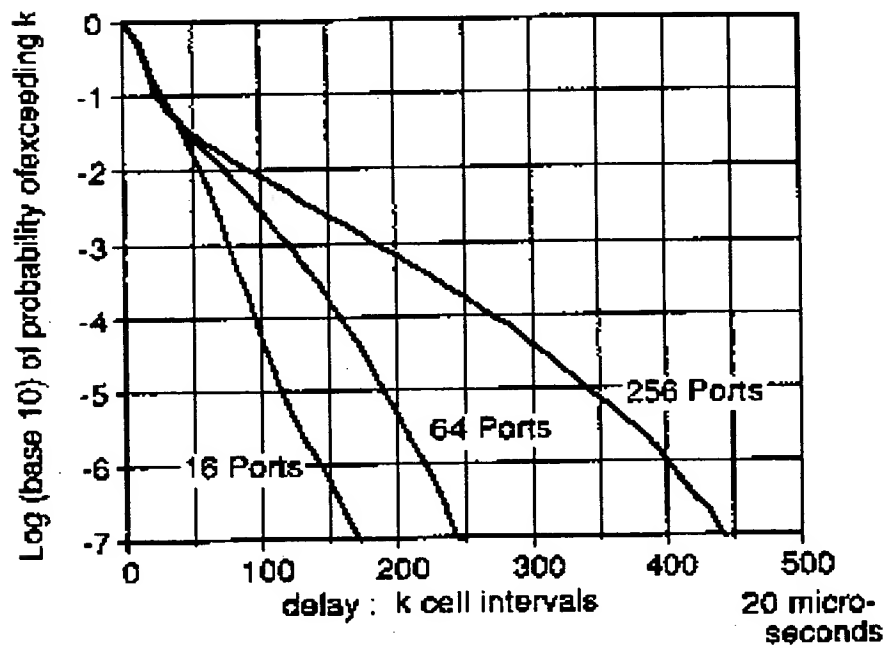


**Fig 16**



**Fig 17**

10/10

**Fig 18****Fig 19**